Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Improve English Pronunciation at Word Level for Thai EFL Learners in Southern Region Using End-to-End Automatic Speech Recognition

Nattapol KRITSUTHIKUL[a*], Kongpop BOONMA[b], Jirapond MUANGPRATHUB[c], Wasan NA CHAI[a] & Thepchai SUPNITHI[a]
[a]*National Electronics and Computer Technology Center (NECTEC), Thailand*
[b]*Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand*
[c]*Faculty of Science and Industrial Technology, Prince of Songkla University, Suratthani Campus, Thailand*
*nattapol.kritsuthikul@{nectec.or.th, gmail.com}

Abstract: ASR (Automatic Speech Recognition) is favorably chosen as a learning technology, which is used for English pronunciation practice. This research aims to build a personalized learning platform to improve English pronunciation at the word level for Thai EFL learners who learn English as a Foreign Language (EFL) by using ASR to detect mispronounced sounds. ASR models are built with an End-to-End learning approach with a Thai-English mispronounced words dataset. The practice of English pronunciation particularly focuses on eleven problematic consonant sounds of Thai EFL students according to the previous studies of English pronunciation in Thai contexts. These eleven consonant sounds are divided into five groups: 1) /ð/-/θ/-/tθ/, 2) /ʒ/-/ʃ/, 3) /dʒ/-/tʃ/, 4) /z/-/s/ and 5) /b/-/p/. The five of Grade 12 Thai Students who are native Thai speakers were selected as sampling process. The result of pre-test and post-test show that the samples have the most problem with the consonant sounds of /ð/-/θ/-/tθ/ (29%), followed by /b/-/p/ (22%), /dʒ/-/tʃ/ (22%), /z/-/s/ (18%) and /ʒ/-/ʃ/ (9%) respectively. In conclusion, this study reveals that 60% of the samples have improved their pronunciation after using our system.

Keywords: English as a Foreign Language (EFL), Mispronunciation, Thai, Automatic Speech Recognition (ASR), End-to-End, Personalized Learning Platform, Computer aided pronunciation training (CAPT)

1. Introduction

Today, English plays a big role as an international language and becomes a dominant language for intercultural communication. One of the most important language skills is speaking skills. Similarly, English is taught as a foreign language in Thailand, but Thai EFL students still have a very low levels of English proficiency (EF EPI, 2021).

The type of English that we speak doesn't matter very much as far as we speak in an intelligible way. If you live in a country where there is no traditional use of English and no people who speak it for general communication purposes, the English pronunciation you are going to speak may reflect the distinction between your native language and English.

Furthermore, the English pronunciation that can be understood in your home country may not be the case in another. Though English is not the media for communication in Thailand, sometimes Thai people use borrowed English words, but pronounced in Thai ways that native speakers may not understand (Wei, 2022).

There are vary of English pronunciation practice software. Most software can detect the right sound, but a few can detect the mispronunciations sound. Hence, this paper aims to design a system as a self-learning material for Thai EFL students to practice their English pronunciation by build an algorithm with the ASR to detect mispronunciation sounds. The ASR is defined as a cutting-edge technology that allows a computer or even a mobile device to identify words that are read aloud or spoken into any sound-recording device. The ultimate purpose of ASR technology is to allow 100% accuracy with all words that are intelligibly spoken by any person regardless of vocabulary size, background noise, or speaker variables. For this research, we use the ASR as a tool to detect mispronunciation sounds. The content in Introduction will be discussed more in Literature review.

2. Literature Review

2.1 English Pronunciation Problems in Thais

English is a pluricentric language; that is to say, varieties of English are emerging with different marked accents. These accents become remarkable in global communication when English plays an important role as a lingua franca. English learners from different parts of the world produce different sounds influenced by their mother tongue, and that possibly causes them a barrier and misunderstanding. Therefore, an international intelligibility relies on the clarity of English pronunciation and utterances, rather than native-like pronunciation (Jenkins, 2002). Several previous studies report problems of English pronunciation found in Thai EFL learners that Thai English learners still have encountered difficulties in mispronunciation of some sounds such as /s/, /z/, /θ/, /r/, /l/, /v/ and /f/ (Moxon, 2021; Sahatsathatsana, 2017; Wei et al., 2002). In addition, Jaroensak & Saraceni (2019) reported that some distinctive features of Thais' English pronunciation found in lingua franca communication were a shift of initial and final consonant sounds. The consonant sound /t/ was shifted into /tʃ/ and the sound /tʃ/ was pronounced as /ʃ/. These mispronunciations possibly cause a barrier to interlocutors' intelligibility; on the other hand, it leads to speakers' loss of confidence and anxiety to speak English.

2.2 Overview of Using a Software as an English Language Learning Material on Thai EFL Learners

Studies on improving English pronunciation skills by using the software have already been widely conducted. The study of Arunsirot (2017) enhanced the English pronunciation competence of Thai students by using Speech Analyzer software. This tool gives the visualizations of the raw waveform and intensity of the sound wave of the voice record of both the student and the native speaker; hence they can compare the contrast in their pronunciations between themselves and the native speaker. As an example, some of the words or sounds are considered to force the sound through the oral cavity differently from the language so the students could know how to control their oral muscles and air pressure from the vocal tract more certainly. The results reported that using the speech analysis software could significantly enhance the students' pronunciation.

Phomprasert (2020) created Detect Me English application, The English correction software to analyze the phonological sounds produced by the students, to create awareness of incorrect pronunciations in Thai Elementary students. The results showed that the sample groups were not aware of many of the English phonological rules.

Parthanasin & Blackford (2015) created a case study of Siri in iPad as voice recognition for English Pronunciation practice. The results of the analysis clearly prove that Siri application could recognize utterances spoken by a native speaker of English better than those spoken by a non-native speaker in all categories. It can be claimed that correct pronunciation is essential for the machine to recognize an utterance.

Graham (2020) used SpeaKIT's speech recognition software to study whether this software can improve English primary school English language learners' skills. The results showed that it is apparent that audio-visual speech recognition has the potential to have something for every child and that it is up to the teacher to identify the student's individual learning style and adapt the implementation of the program according to the needs of the students in their specific classroom.

2.3 End-to-end Speech Recognition

The speech research community found a reliable and effective approach (Jinyu, 2022) both in lab prototype (Watanabe et al., 2017) and industry prototype (Li et al., 2020). The major advantages of End-to-End approach are reduced intensive workload of the audio engineer robustness acoustic models. By comparison with the traditional speech system based on Hidden Markov Models (HMMs), End-to-End approach required a few inputs e.g. (1) raw audio and (2) transcription of the raw audio input, and so on. One disadvantage of End-to-End approach is that they require high performance computing power especially GPUs at each step.

End-to-End ASR relies on sophisticated pipelines composed of multiple algorithms and fine-tuning processes by the audio engineer. To achieve the ASR tasks, an End-to-End ASR pipeline was constructed from various concepts and tutorials e.g., Very Deep Learning and M5 Network (Wei et al., 2016), Deep Speech (Hannun et al., 2014), Speech Command Classification with TorchAudio (Gatignol, 2022), and Thai Speech Command Recognition with TorchAudio (“Thai Speech Command Recognition with TorchAudio”, 2021). The pipeline allows us to apply the ASR as the mispronunciation detection tool. In overall, the pipeline applied very deep convolutional neural networks (CNNs), up to 34 weight layers, to processing the raw audio data as inputs. The M5 network architecture is used to filter the raw data into the receptive fields around 20 ms each. This size is similar to speech processing applications that often use receptive fields ranging from 20 ms to 40 ms for training and testing the network. Finally, we construct the model by applying the pipeline in the training process based on the Thai mispronunciation corpus.

3. System Architecture

In this section, we will describe the proposed system and system overview for our proposed system, an E-learning platform.

3.1 System Requirements and Design

The proposed system aims at improving English pronunciation for Thais by detecting the mispronounced sounds from the learner and giving the correct pronunciation and advice to them. By doing that, the learner would know what sounds they speak wrong so they could correct those wrong sounds to improve their English pronunciation. Figure 1 shows the system architecture that outlined each major component for the proposed system.

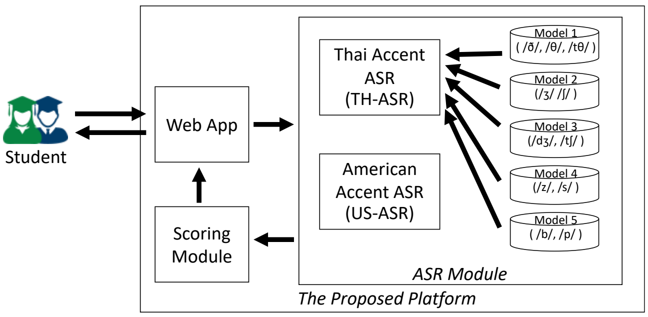

Figure 1. System Architecture

The system is developed as a web-based application using HTML5, CSS3, ECMAScript 2020 as front-end framework, Flask, a Python micro web framework, as back-end framework in advantage of convenient of Python libraries invocation, and MySQL as a relational database management system (DBMS) to store the system information, e.g. learning activities, scores, and so on. The system consists of two modules which are *ASR Module* and *Scoring Module*.

ASR Module is used to predict voice audio from the student to the system by either uploading voice audio files or directly streaming audio into the module. There are two types of ASR, American Accent ASR powered by Google speech-to-text API, and ASR for five mispronunciation models as mentioned in the Model Training Process.

Scoring Module is used to manage the student learning path based on the student performance interaction in the system. Figure 2 shows the scoring process, started by the user have to speak the word from the system, the system then detects the user's speech with Google speech-to-text API to match the user's speech with the word user speaks. If the Google speech-to-text API output matches with the word user speaks, we consider that user has no problem with that sound, but if Google speech-to-text API output is not equal to the spoken word.

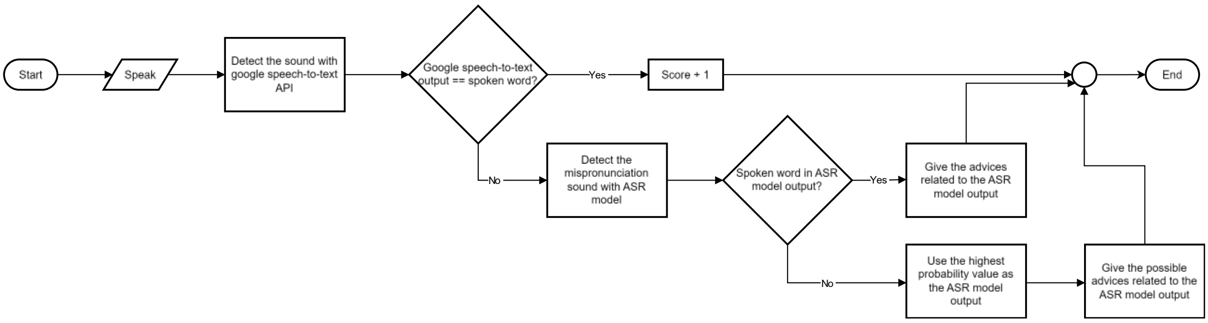

Figure 2. Scoring Process

The system is going to detect mispronounced sounds with 5 ASR models. If ASR predicted word output equal to spoken word, the output is sent to the user to suggest the correct pronounced sounds. But if ASR predicted word output is not equal to spoken word, the output is going to be the highest probability value from 5 ASR models. For the process above, the process can allocate the data adequately for indicating the student's mispronouncing. The system will generate suggestions to the student. Finally, both ASR Module and Scoring Module store the processing results into the database.

3.2 Audio Acquisition

From our study, the sounds that have problems in English pronunciation for Thais have come up with 11 sounds, namely /ð/, /θ/, /tθ/, /ʒ/ /ʃ/, /dʒ/, /tʃ/, /z/, /s/, /b/ and /p/. We record those sounds by native Thai students, undergraduates, and teaching professionals in English and non-English subjects. The record is created for Thai-English mispronounced pronunciation corpus to train the ASR model to detect mispronounced sounds. The corpus (1,874 sounds) is divided into 5 dataset 1) /ð/, /θ/, /tθ/ (370 sounds), 2) /ʒ/ /ʃ/ (370 sounds), 3) /dʒ/, /tʃ/ (368 sounds), 4) /z/, /s/ (380 sounds) and 5) /b/, /p/ (386 sounds).

3.3 Model Training Process

To build the model, TorchAudio (Yang et al., 2022), a library for audio and signal processing with PyTorch, is introduced for building ASR models to detect mispronounced sounds. The audio data have to be set in 2-channel audio. The training data file consists of 2 classes 1) audio filename and 2) transcript for each audio in Comma-separated values (CSV) file format. The model gives an output consisting of 2 data. 1) a predicted word and 2) a confident value. The predicted word is the word that the model predicts from the voice audio. The confidence value is the probability of the predicted word. The output will be used as a source to indicate whether the student pronounces this word correctly or not.

At the end of this process, the five ASR models are conducted according to the five datasets that are described in the Audio Acquisition section.

4. Experiment

We test the system with 5 native Thai speaker students studied in PSU Witthayanusorn Suratthani School. Testers have to test with our learning process starting with pre-test. After finishing the pre-test, testers have to learn the lessons order by lesson recommendation. Finally, Testers have to do the post-test to see the differences after using the system. Post-test words are all similar to pre-test words.

The learning starts on the pre-test. The pre-test has 15 words divided into 5 groups according to datasets. There are 3 words for each group as shown in Table 1. After finishing the pre-test, the system prepares the lesson according to the user's score.

Table 1. *Group of words in for pronunciation practice*

Group	Word
/ð/-/θ/-/tθ/	altogether
	smooth
	though
/ʒ/-/ʃ/	bashful
	chicanery
	distinguish
/dʒ/-/tʃ/	charge
	fragile
	indulge
/z/-/s/	decease
	overseas
	practice
/b/-/p/	belonging
	bilingual
	dabble

5. Results

The three students (60%) improved after using our system. Nevertheless, the highest score a learner could do is only 5 out of 15 points.

According to data on the training datasets, the ASR models by /ð/-/θ/-/tθ/, /ʒ/-/ʃ/, /dʒ/-/tʃ/, /z/-/s/ and /b/-/p/ have a Word error rate (WER) of 44%, 60%, 56%, 29%, 33% respectively. We can see that all models have low accuracy. The reason why all models have low accuracy is that our corpus does not have enough amount of data to train the appropriate accuracy ASR model.

We also found that there are 7 words that none of the samples could speak correctly which are "altogether", "smooth", "though", "fragile", "decease", "belonging", and "dabble".

The experimental result shows that samples have problem with /ð/-/θ/-/tθ/ the most (29%) followed by /b/-/p/ (22%), /dʒ/-/tʃ/ (22%), /z/-/s/ (18%), /ʒ/-/ʃ/ (9%) according to Google speech-to-text outputs.

The percentage of mispronunciation output from the ASR models. The highest mispronunciation output count is ʒ/-/ʃ/ (61%) followed by /b/-/p/ (11%), /z/-/s/ (11%), /ð/-/θ/-/tθ/ (10%) and /dʒ/-/tʃ/ (7%). However, this data is still unreliable because low accuracy models provide many wrong outputs.

6. Conclusion, Discussion, and Future Works

In this paper, we proposed a prototype of a personalized learning platform to improve English pronunciation at the word level by using ASR to detect mispronounced sounds and give correct pronunciation. Our study aims at 11 consonant sounds that Thai EFL learners have problems with (see Table 1)

Based on the experimental results, it can be concluded that using ASR to detect mispronounced sounds has a potential to improve pronunciation skills. Moreover, we also found that samples have problems with /ð/-/θ/-/tθ/ the most (29%) followed by /b/-/p/ (22%), /dʒ/-/tʃ/ (22%), /z/-/s/ (18%), /ʒ/-/ʃ/ (9%).

Since the accuracy of the prototype depends on the ASRs, we intend to employ better ASR models. We also intend to find more detection method of mispronunciation for Thais in future work. Furthermore, samples and the amount of words on the experiment are still too low to provide reliable results. Therefore, in further research, we want to apply more samples and words to extend experimental data for reliable results. The corpus also needs to expand to provide proper accuracy. We will build the model by applying a multi-class classification training approach so that we can merge the five ASR models into a single model.

Acknowledgements

We would like to thank Tirapond Jaroensak, Phongsatorn Ousakulwattana and the other 16 people for helping on collect the data for the Thai-English accent corpus and Tankwan Khongcharoen for helping.

References

Alexis Gatignol. (2022). "Speech Command Classification with TorchAudio," https://github.com/pytorch/tutorials/blob/master/intermediate_source/speech_command_classification_with_torchaudio_tutorial.py, 2022.

Arunsirot, S. (2017). "Implementing a Speech Analyzer Software to Enhance English Pronunciation Competence of Thai Students", 2017.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. (2014). "Deep speech: Scaling up end-to-end speech recognition," arXiv preprint arXiv:1412.5567, 2014.

EF EPI (2021). "EF English Proficiency Index", https://www.ef.com/wwen/epi/

Graham, S. (2020). "An Exploratory Case Study to Investigate Perceived Pronunciation Errors in Thai Primary School Students Using Audio-Visual Speech Recognition, Teaching English with Technology", 2020.

J. Li, R. Zhao, Z. Meng, Y. Liu, W. Wei, S. Parthasarathy, V. Mazalov, Z. Wang, L. He, S. Zhao, et al. (2020). "Developing RNN-T Models Surpassing High-Performance Hybrid Models with Customization Capability," in Proc. Interspeech, 2020, 3590–4.

Jaroensak, T., & Saraceni, M. (2019). ELF inThailand: Variants and Coinage in Spoken ELF in Tourism Encounters. rEFLections, 26(1), 115–133. https://doi.org/10.61508/refl.v26i1.203948

Jenkins, J. (2002). "A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language", Applied linguistics, 23(1), 2002, 83-103.

Jinyu Li. (2022). "Recent Advances in End-to-End Automatic Speech Recognition," arXiv preprint arXiv: 2111.01690v2, 2022.

Moxon, S. (2021). "Exploring the Effects of Automated Pronunciation Evaluation on L2 Students in Thailand", IAFOR Journal of Education, 9(3), 2021, 41-56.

Pathanasin, S., & Blackford, M. (2015). "Voice Recognition for English Pronunciation Practice: A Case Study of Siri in iPad", 2015.

Phomprasert, J. (2020). "Creating Awareness of Incorrect English Pronunciation in Thai Elementary School by using the Detect Me English Application", 2020.

S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi (2017). "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," IEEE Journal of Selected Topics in Signal Processing, 11(8), 2017, 1240–53.

Sahatsathatsana, S. (2017). "Pronunciation problems of Thai students learning English phonetics: A case study at Kalasin University", Journal of Education, Mahasarakham University, 11(4), 2017.

Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. (2016). "Very Deep Convolutional Neural Networks for Raw Waveforms," arXiv preprint arXiv:1610.00087, 2016.

Wei, Y. (2002). "Insights into English Pronunciation Problems of Thai Students".

Wei, Y., & Zhou, Y. (2002). "Insights into English Pronunciation Problems of Thai Students", 2002.

Workshop on NLP/AI R&D in iSAI-NLP-AIoT 2021 (2021). "Thai Speech Command Recognition with TorchAudio," https://colab.research.google.com/drive/1ensKfWzt6WEvmAZTrtMtyUrX1i5JBkMk#scrollTo=yUHwDMnYqI8l, 2021.

Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg1, Edward Z. Yang, Jason Liany, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélairy, and Yangyang Shi. (2022) "TorchAudio: Building Blocks for Audio and Speech Processing," arXiv preprint arXiv:2110.15018v2, 2022.